ERDC/CERL TR-09-15

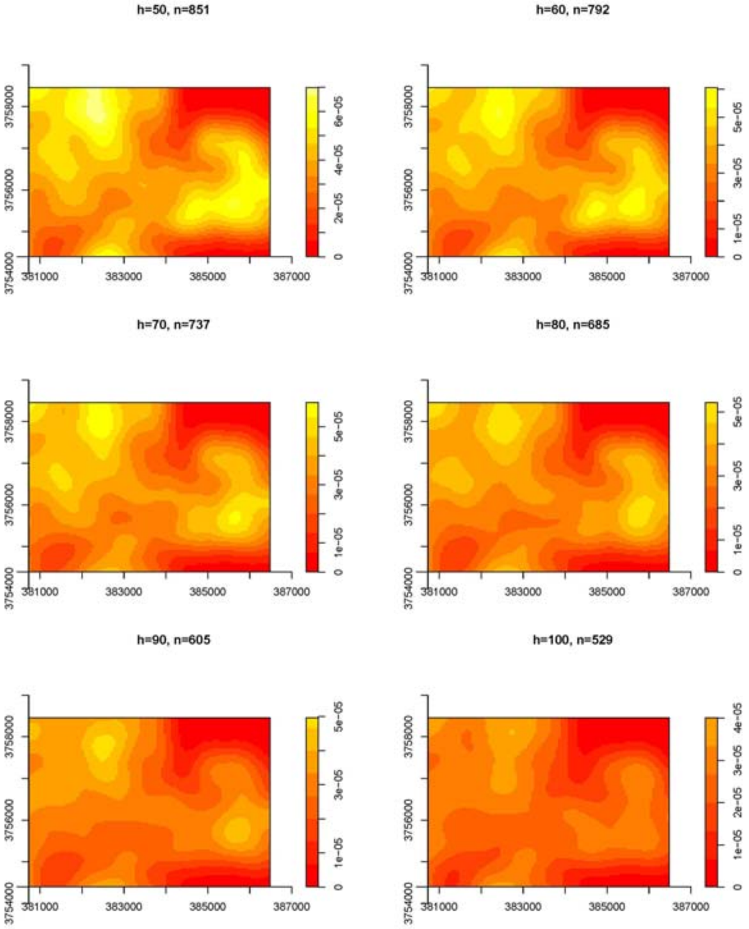
US Army Corps
of Engineers®
Engineer Research and
Development Center

Actionable Cultural Understanding for Support to Tactical Operations (ACUSTO)

The Effect of Data Quality on Spatial Analysis Results

Luis Galvis, Patrick J. Guertin, and William D. Meyer

June 2009



Construction Engineering
Research Laboratory

ERDC/CERL TR-09-15
June 2009

Actionable Cultural Understanding for Support to Tactical Operations (ACUSTO)

The Effect of Data Quality on Spatial Analysis Results

Luis Galvis

University of Illinois
Champaign, IL 61820

Patrick J. Guertin and William Meyer

Construction Engineering Research Laboratory (CERL)
U.S. Army Engineer Research and Development Center
2902 Newmark Dr.
Champaign, IL 61822-1076

Final Report

Approved for public release; distribution is unlimited.

Prepared for Headquarters, U.S. Army Corps of Engineers
Washington, DC 20314-1000

Under 21 2040

Abstract: Developing cultural information into cultural knowledge for military operations is predominantly an intelligence activity that takes place within the Military Decision Making Process. The products of such efforts are routinely classified and unusable by the tactical war fighter. The Actionable Cultural Understanding for Support to Tactical Operations (ACUSTO) research effort was undertaken to provide a product for enhanced cultural understanding that will be accessible to the tactical war fighter. This is done by combining spatial and explicit content analysis of open source news media to provide cultural understanding in the operational environment that can be disseminated down to the lowest tactical level. The development of actionable intelligence for counterinsurgency parallels the study of civilian criminal events (widely covered in open source media) and can exploit the methodological approaches that emphasize spatially explicit information. Crime research is conducted at aggregate levels, which implies the aggregation of a series of points representing events to areas representing higher scales. This work focused on data quality in point pattern analysis, and on the effect that different levels of data accuracy and precision have on policy recommendations.

DISCLAIMER: The contents of this report are not to be used for advertising, publication, or promotional purposes. Citation of trade names does not constitute an official endorsement or approval of the use of such commercial products. All product names and trademarks cited are the property of their respective owners. The findings of this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

DESTROY THIS REPORT WHEN NO LONGER NEEDED. DO NOT RETURN IT TO THE ORIGINATOR.

Table of Contents

List of Figures

Preface

This study was conducted for the Assistant Secretary of the Army for Acquisition, Logistics, and Technology (ASAALT) under Project 62784AT41, “Military Facilities Engineering Technology,” Work Unit 21 2040, “Social-Cultural and Environmental Data Fusion Models.”

The work was completed under the direction of the Ecological Process Branch (CN-N) of the Installations Division (CN), Construction Engineering Research Laboratory (CERL). The CERL Project Manager was William D. Meyer. Luis A. Galvis is affiliated with Arizona State University, Tempe, AZ. Alan B. Anderson is Chief, CN-N, and Dr. John T. Bandy is Chief, CN. The associated Technical Director was Dr. William D. Severinghaus. The Director of CERL is Dr. Ilker Adiguzel.

CERL is an element of the U.S. Army Engineer Research and Development Center (ERDC), U.S. Army Corps of Engineers. The Commander and Executive Director of ERDC is COL Gary E. Johnston, and the Director of ERDC is Dr. James R. Houston.

1 Introduction

Background

Developing cultural information into cultural knowledge for military operations is predominantly an intelligence activity that takes place within the Military Decision Making Process (MDMP). MDMP includes mission analysis that produces an intelligence assessment, evaluation of courses of action, and re-evaluation of intelligence assessment. Intelligence Preparation of the Battlefield (IPB) is performed before, during, and after the mission analysis phase of the MDMP. Recent Army field manuals and lessons learned documents emphasize the role of Every Soldier as Sensor (ES2) in providing information for IPB. The incorporation of cultural knowledge into IPB is recognized as especially critical for planning and implementing counterinsurgency operations. In practice, IPB involves collecting data manually or through sensors coupled with computer analysis by highly trained intelligence analysts. The products produced from these efforts are routinely classified and subsequently unusable by the tactical war fighter operating at the brigade combat team level.

The Actionable Cultural Understanding for Support to Tactical Operations (ACUSTO) research effort was undertaken to provide a product for enhanced cultural understanding that will be accessible to the tactical war fighter. This is accomplished through a combination of spatial and explicit content analysis of open source news media to provide cultural understanding in the operational environment (OE) that can be disseminated down to the lowest tactical level. This work approaches the development of actionable intelligence for counterinsurgency by drawing parallels with the study of civilian criminal events, such as homicides, vehicle thefts, and gang violence, and by exploiting the methodological approaches that emphasize spatially explicit information. This *spatial analysis of crime* (Anselin et al. 2000, Messner and Anselin 2004) builds on the well-established methods of spatial data analysis and spatial statistics, and applies these in the context of criminal events that occur at specific locations.

Criminology theories point to the chief role that space and place has on understanding why crime events occur. By and large, research on crime is conducted at aggregate levels such as neighborhoods, census tracts, cities and so on. This usually implies the aggregation of a series of points representing events to areas representing higher scales.

A case study in Watts, Los Angeles, CA forms the study area. When geocoding crime data, problems emerge when locational information cannot be seamlessly translated into its companion coordinates. This larger mismatch is not randomly distributed; areas that are newly developed, economically deprived, or located at the city outskirts, are more likely to have incomplete geographic data, lack detailed disaggregated information, or to be associated with outdated information. This is a critical issue as homicides or other type of crime events or events significant to the military are not homogeneously distributed across neighborhoods or space in general. Instead, such events are clearly associated with structural conditions such as social and economic stratification or demographic characteristics.

This work evaluates the extent to which this process introduces certain biases due to the imprecision of the location of the events, which in turn may result in erroneous strategies to fight crime or in the incorrect selection of a course of action (COA) in a military context, which could result in unintended consequences and an inefficient use of resources. Central tendency measures and second-order statistics are evaluated in different scenarios of homogeneous and heterogeneous underreporting levels, and different precision levels to evaluate their incidence on the clustering and cluster detection statistics.

More specifically, this work focuses on the issue of data quality in point pattern analysis, and on the effect that different levels of data accuracy and precision have on policy recommendations.

Objectives

The objective of this stage of research was to characterize data quality in point pattern analysis for their potential application to the ACUSTO research effort.

Approach

1. A method for point pattern analysis was determined, and analytical tools were defined.
2. A dataset was determined to test the method (homicides in Watts, Southeast Los Angeles, CA).
3. Data quality was assessed, and likely compromises to data quality were defined as under-reporting and imprecision.
4. Point pattern analyses were simulated for data of varying quality, and conclusions were drawn regarding the results of those analyses.

Mode of technology transfer

This report will be made accessible through the World Wide Web (WWW) at URL: http://www.cecer.army.mil

2 Applied Methods

Actionable context of data quality in GIS based intelligence applications

GIS-based intelligence support has value added function to military sustainability operations. Applications including terrain evaluation, logistics management, and intelligence analysis (Satyanarayana and Yogendran 2009), including socio-cultural factors are all with in the realm of GIS. These applications are varied and range in ability to support operational, strategic, and tactical goals of the warfighter.

Examples of GIS analysis in support of Sustainability Operations include use by the Multi-National Brigade during operations in Bosnia in 2003. Applications were used for a myriad of support functions including: (1) tracking refugees and analyzing logistical support, (2) analyzing patterns in weapon caches collections, and (3) identifying possible threats to peace (Reichman 2008). GIS-based intelligence analysis was further expanded during the current conflicts in Iraq and Afghanistan where it provided the opportunity to be exploited for dealing with crime, terrorism, and force protection (Defenselink). Many of the applications used for these functions are offshoots of civilian applications developed for intelligence-led policing in many of America's large cities.

In these civilian policing applications, GIS is used to identify strategic necessities, tactical hotspots, and socio/cultural-linkages to better plan and allocate resources (Ratcliffe 2004, Bichler-Robertson and Johnson 2001). Applications are targeted across a broad hierarchy to meet strategic (i.e., budgeting, manpower) issues as well as assessing areas of increased risk for tactical deployment of limited policing resources.

Military and civilian use of spatial data analysis to support intelligence-based resource allocation is only effective if used within the parameters of the available data. Quality results are dependent on quality data inputs. Many factors are important to "data quality" as applied to the analysis methods. In general, factors include:

1. Quality Assurance/Quality Control: Uniform data collection/storage methodologies etc.

2. Sample/ Size: Adequate sampling to cover the populations/scenario of interest, including ample size to meet statistical concerns and appropriate representation of stratum
3. Focused Sampling Strategy: Design including sampling parameters adequate in explaining the phenomenon of interest.

It is in the formal terms and concepts of the methods used (i.e., point pattern analysis) within this report that the quality of data is tested to showcase the importance of data quality on actionable results. As oftentimes it is easier to conduct a comprehensive analysis using unclassified civilian data sources, the analysis is performed on crime data from Watts County, CA. This ensures that the widest possible dataset is used while maintaining a direct link to military application of methodologies.

Point pattern analysis

A primary goal in the study of events that conform a point pattern is the identification of a spatial arrangement of events suggesting that the process analyzed goes beyond complete spatial randomness (CSR), for which there is simply not much to say other than that the events are equally likely to occur at any location (Waller and Gotway 2004). The analysis of point patterns comprises a set of tools to study those deviations from CSR, specifically the clustering and the clusters are of special interest. A process that exhibits clustering is interesting because it suggests that the locations where events take place has an underlying structure.

Central tendency measures: Intensity

In statistics, the measures of central tendency ordinarily provide a good starting point to describe the general characteristics of the sample studied. The sample mean is employed to account for that central tendency since this is an unbiased estimator of the expected value of the random variable analyzed. In spatial processes, this statistic is represented by the *intensity* of the point process. For a sample of points X that is known to be homogeneous and that takes place over a two-dimensional space W, the intensity statistic is calculated as:

$$\bar{\lambda} = \frac{n(X)}{area(W)} \qquad \text{Eq 1}$$

This statistic represents the average number of points per area. One important issue to be considered is how the study area is bounded. Usually it is defined in three different ways:

- as the smallest encompassing administrative unit in which the sample points were observed
- as the bounding box equivalent to the rectangle bounding the minimum and maximum *x* and *y* coordinates
- as the *convex hull,* which is defined as the tightest polygon that surrounds the points.

In practice, this is defined rather arbitrarily disregarding the fact that point pattern analysis can be greatly affected by the choice made about the boundaries of the study area, which will be explored in more detail later. The choice of the study area has also been shown to affect the results of the clustering tests (Waller and Gotway 2004, pp 141–146).

Global clustering statistics

The methods of analysis of clusters and clustering in crime studies are of paramount importance, as they provide a formal framework for police and security enforcement agencies to direct limited resources to be spent in operations to critical time periods and places i.e., "hot spots." Likewise, agencies can save resources by reducing operations or surveillance where criminal activities are not significantly high, in what has been called "cold spots." This section summarizes some of the most used statistics and statistical devices to study clustering and to detect clusters of unusual (high or low) occurrence of crimes.

Point level clustering

K Function

Ripley introduced the *K* function as a statistical tool to analyze the second order moment of a point pattern process (Ripley 1976). Its usefulness comes into play when there is need for the distinction between a CSR, a regular, or a clustered point pattern process. For a CSR that is used as a benchmark process, $K(h)= \pi h2$. What is more important in a point pattern process is the deviations from that benchmark $K(h) < \pi h2$ indicates a "regular" point pattern process, and when $K(h) > \pi h2$ the process is said to be "clustered." The three simulated point patterns in Figure 1 illustrate these three processes.

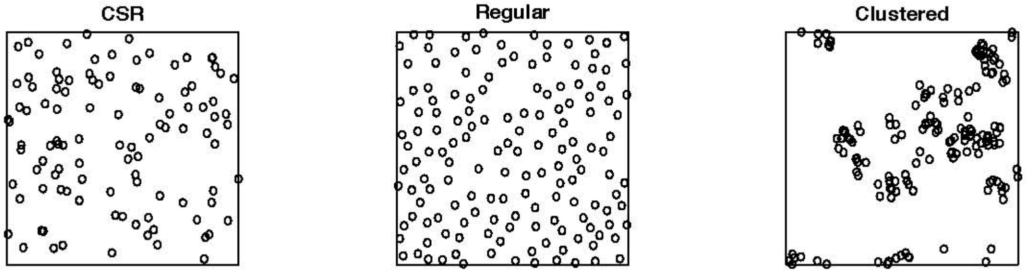


Figure 1. Simulated point patterns.

Mathematically the *K* function can be expressed as:

$$K(h) = \frac{E(N_0(h))}{\lambda} \quad \text{Eq 2}$$

Equation (2) can be understood as the ratio of two components: in the numerator we have the expected number of further events lying within a distance *h* from an arbitrary event of the process, and the denominator is just the intensity of the process, λ, that comes from Equation (1).

The best way to appreciate the meaning of the *K* function is to plot the values of K(h) against *h* as it is shown in Figure 2.

K function inference

To assess statistical inference on the nature of the process, first calculate the envelopes that result from simulating a series of random processes, to later obtain a confidence interval with the highest and lowest values for the K(h) for different values of h. The purpose of the envelopes is to simulate the boundaries of the region within which K(h) is statistically equal to πh^2.

The darker line that corresponds to the observed K(h) (shown in Figure 3) lies within the confidence interval for a CSR, below the confidence interval for a regular pattern, and above the confidence interval for a clustered pattern.

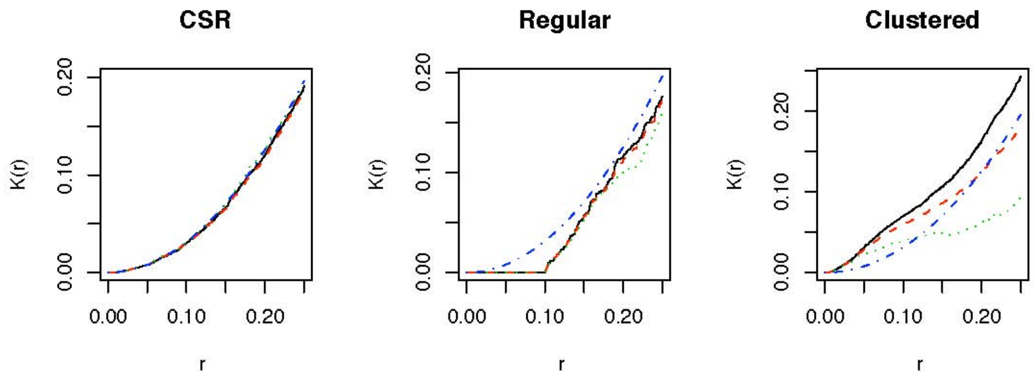


Figure 2. *K* function.

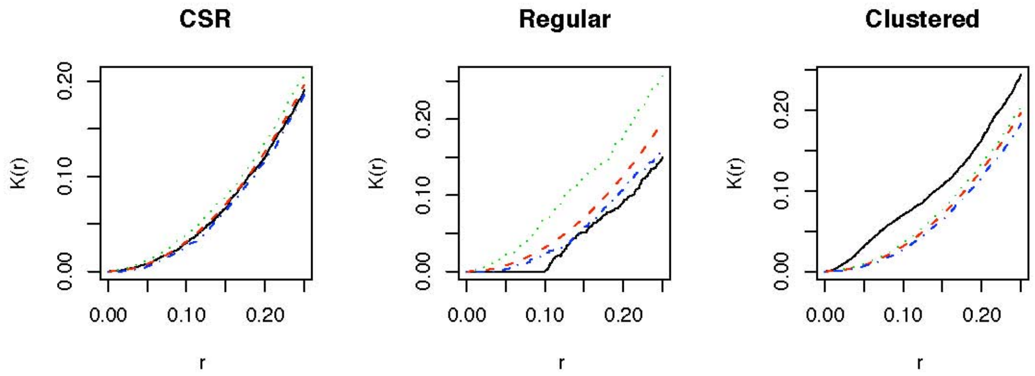


Figure 3. *K* function envelopes.

Aggregate measures: Global Moran's I

Moran's I is a statistic used to measure spatial auto-correlation at a global level i.e., to have an indicator of clustering. In essence, it is a cross product statistic that is a special case of the gamma statistic or the general cross product statistic. The latter is used in spatial statistics to show the match between locational similarity and value similarity. In terms of value similarity, note that Moran's I is similar in essence to the Pearson correlation statistic. One thing to observe is that the Pearson statistic gives a bivariate correlation, whereas the Moran's I is used to calculate univariate correlation (correlation of one variable with itself), and for that reason is called an "autocorrelation statistic." To better explain, the following expression for Moran's I shows what is meant by autocorrelation in a cross product statistic. Let :

$$Z_i = x_i - \mu$$

and

$$S0 = i_j W_{ij}$$

where:

x_i is the variable of interest
μ is its sample mean.

Then Moran's I is given by:

$$I = \left(\frac{N}{S_0}\right)\frac{\sum_i \sum_j W_{ij}.Z_i Z_j}{\sum_i Z_i^2} \qquad \text{Eq 3}$$

In Equation(3), it is shown that the cross product originates in the product of the variable *Z* at location *i* and *Z* at location *j*. For that reason, Moran's I is an autocorrelation statistic that is a special case of the general gamma statistic.

Note that, when W_{ij} is row-standardized as a result of dividing each row by the row sum, the sum of all the elements of each row become one and the term S_0 is equal to N. In this case, Equation (3) can be simplified to:

$$I = \frac{\sum_i \sum_j W_{ij}.Z_i Z_j}{\sum_i Z_i^2} \qquad \text{Eq 4}$$

Local clustering statistics

When studying local-specific statistics, the focus of analysis is no longer in a single statistic that summarizes the global pattern, but rather in individual statistics that identify the places where there is a deviation from a pattern found by chance. For events that are recorded at a point level. one can evaluate the presence of clusters with measures such as the Spatial and Temporal Analysis of Crime (STAC) algorithm, while for areal data, there is (among others) a local version of the Moran's I (which is briefly introduced in the following paragraphs).

Point level clusters

At the point level, one of the most common methods used for detection of clusters in crime analysis is the STAC algorithm (Williamson et al. 2001). The algorithm was first developed by Bates (1987) and enhanced later by Block (1995), and has been made available under CrimeStat, a free package for crime analysis (Levine 2005).

In general terms, the algorithm that runs under CrimeStat overlays a grid structure, the size of which the user can customize. STAC then makes center on every node of the grid to draw the circles that become the search area for events. The events lying within each circle are counted and the circles are ranked according to the number of points. Because there are overlapping circles, if there are points belonging to two circles or more, these circles are merged and the algorithm keeps on ranking them until there is no more overlapping. The result from this procedure yields what has been termed the Hot Clusters, which are used to find the Hot Spot Areas by means of a convex hull fitted to the points in the Hot Clusters. Levine (2004) provides specific details on the algorithm and on other tools for crime analysis.

Aggregate measures: Local Moran's I

Moran's I can be thought of as a summary of the local Moran's I, I_i, that belongs to a general set of statistics named Local Indicators of Spatial Association, LISA (Anselin 1995). Note that, from Equation (4) it is possible to obtain the following expression for I_i:

$$I = \left(\frac{Z_i}{m_2} \right) \sum_j W_{ij} Z_i Z_j \qquad \text{Eq 5}$$

where $m_2 = {}_i z_i^2$

Going back to Equation (4), it follows that:

$$I = {}_{\underline{N}},$$

i.e., the global Moran's I is an average of the local Moran I_i. One of the distinctive features of the local indexes is the ability to detect local clusters. Moran's I in particular, allows the differentiation of high intensity clusters, low intensity clusters, and spatial outliers. High intensity clusters, also called High-High clusters, are those areas with a high incidence of events that are surrounded by zones where the occurrence of the event of interest is equally higher. Conversely, Low-Low clusters point to areas displaying a low incidence of events in a vicinity where events occur with similar low intensity. On the other hand, spatial outliers correspond to Low-High and High-Low spots, which are of special interest as they represent locations where there is an unusual intensity of the events that is not shared by the surrounding locations.

3 Data: Homicides in Watts, Southeast Los Angeles, CA

Data Description

Figure 4 shows the context of Watts, a populated place in southeast Los Angeles, CA. As of the 2000 Census, the total population living in the district was 22,847 within an area of 9.27 squared miles, that yields a density of 2,464 inhabitants per square mile. Figure 5 shows a point pattern dataset representing the location of homicides in Watts for the period 1980-2000.

The following empirical analysis seeks to illustrate what has been discussed so far. As these data represent a realization of a point pattern process taking place over space, this case study resembles other events found in crime data analyses, and processes of interest for agencies dealing with crime incidents.

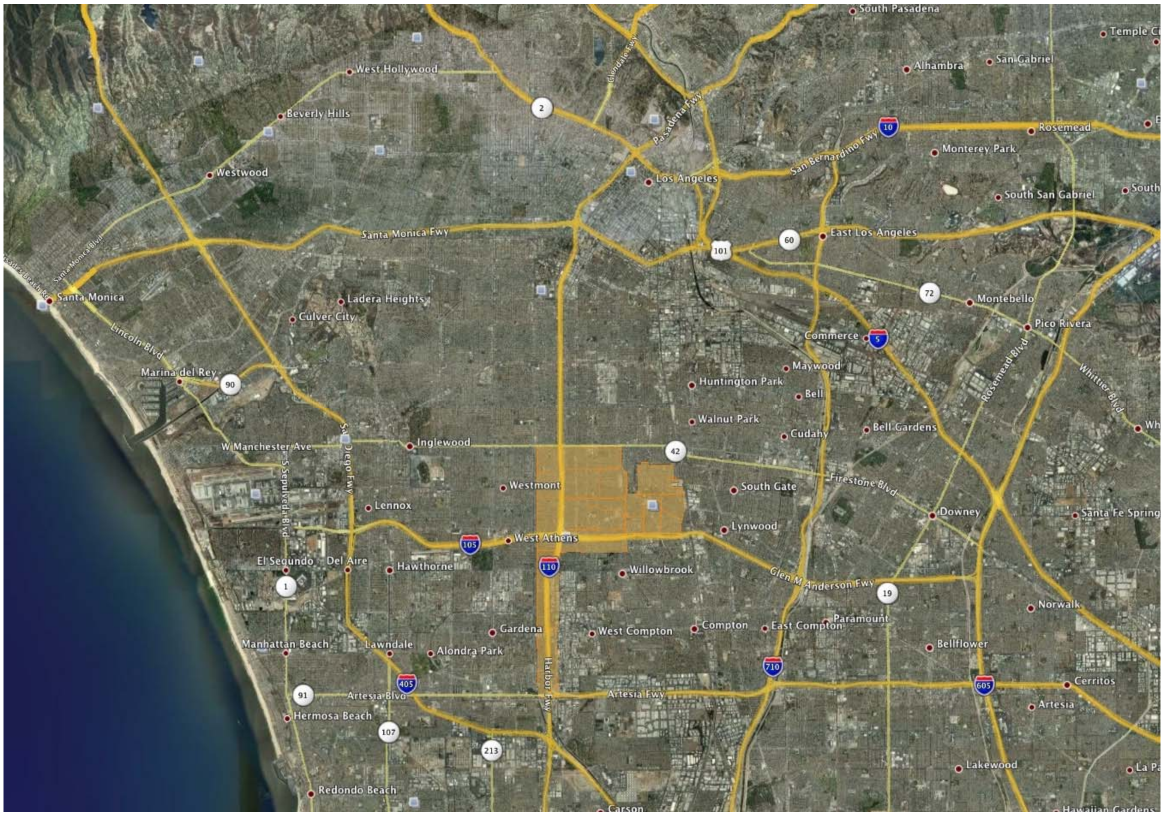


Figure 4. Geographic context of Watts, Los Angeles, CA

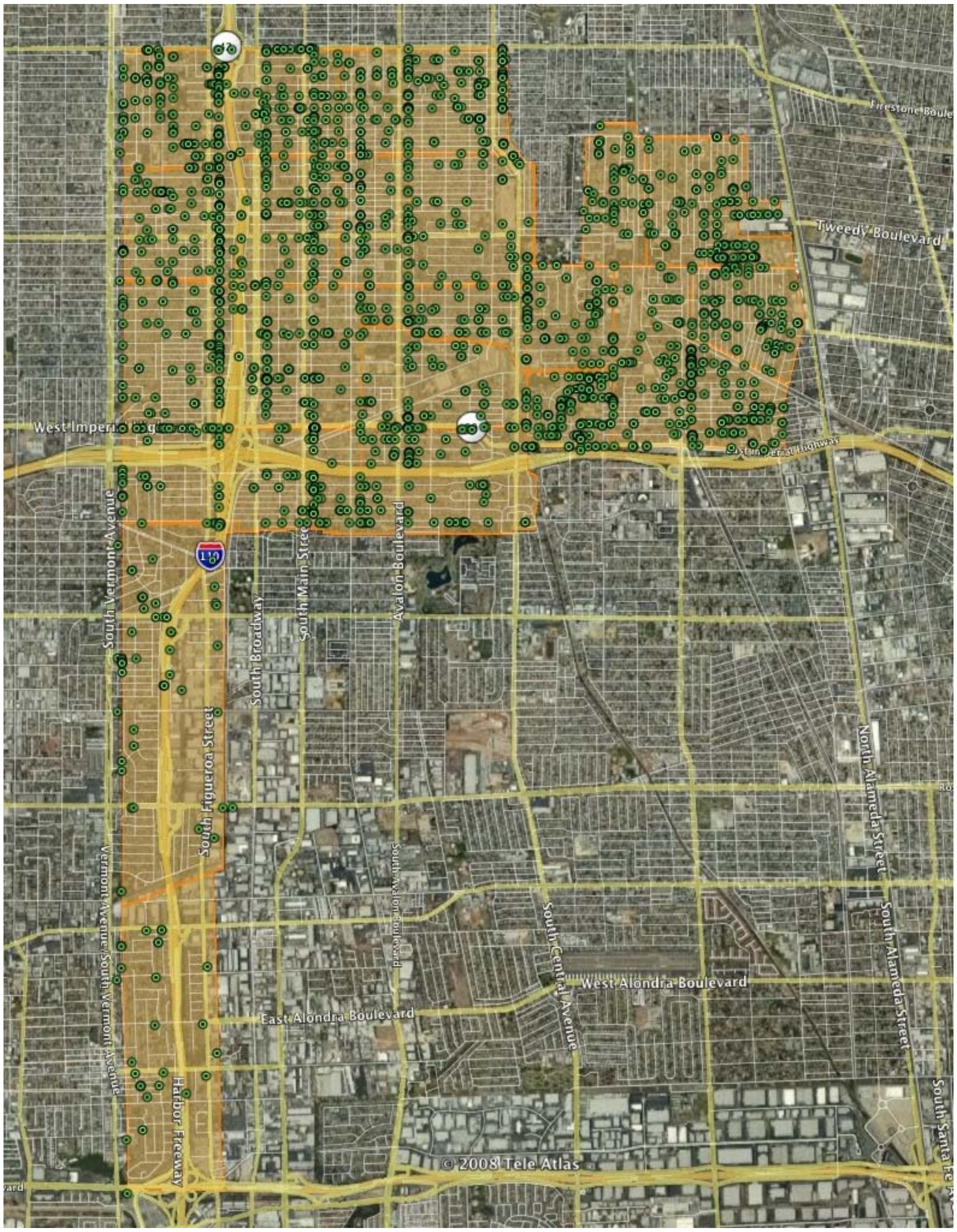


Figure 5. Homicide events in Watts, Los Angeles, CA.

Spatial analysis of homicide data in Watts

This section presents a general description of the data used for the analysis and the results of the statistic tests on global and local spatial autocorrelation, first by looking at the events at point data level, then by moving to areal data by aggregating the events by census blocks. The following sections use just the northern part of Watts to compact the study area, and also to accommodate the fact that the southern strip accounts for only a few incidents (Figure 5). Figure 6 shows the new background map and the homicide events.

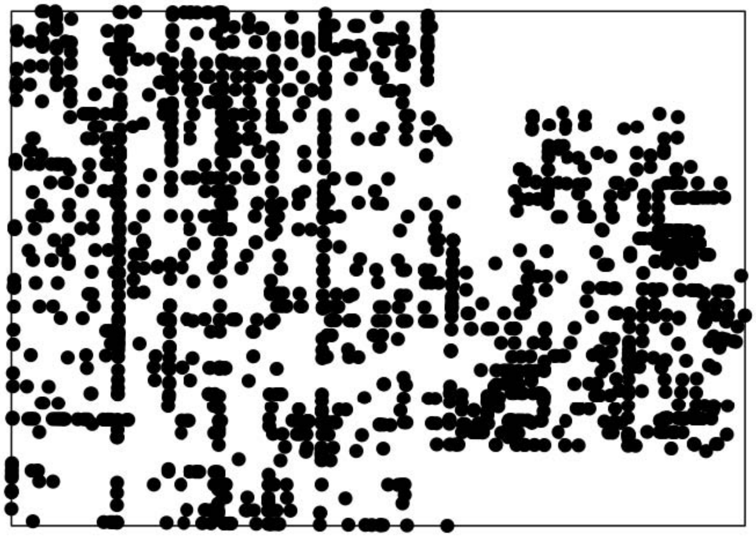


Figure 6. Point pattern of homicide events in Watts.

At point data level

Kernel densities

Kernel densities were estimated using the program *Spatstat*, which runs under *R* (Baddeley and Turner 2005). Kernel densities show a nonparametric estimation of the probability density of a random variable, which, in this case, is the location of events in space. Kernel densities allow us to identify the peaks of the distribution and their respective location to have an indication of where the majority of events occur or where they are concentrated.

Figure 7 shows three core areas where the majority of points are concentrated. As will be shown later in the study of local clusters, the core of the local clusters are found to match the location of the peaks in the kernel density. Of course, this may be influenced by the bandwidth selected to estimate the kernel density. Larger bandwidths are known to result in smoother surfaces that may not reflect correctly the location of local clusters. Conversely, smaller bandwidths will result in spiky maps that will then show a great number of peaks of the distributions that may not correspond to clusters.

The selection of the bandwidth is sometimes subjective; this may be a source of unreliability in the conclusions. Nonetheless, the use of kernels allows us to visually simplify a given plot of point or event locations that in itself may be overloaded; it may be argued that Figures 6 and 7 show this.

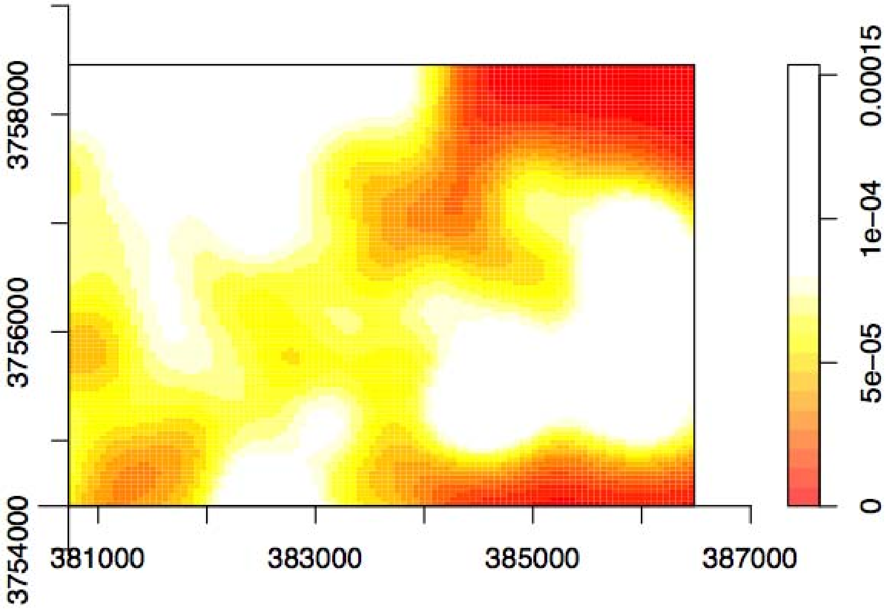


Figure 7. Kernel density.

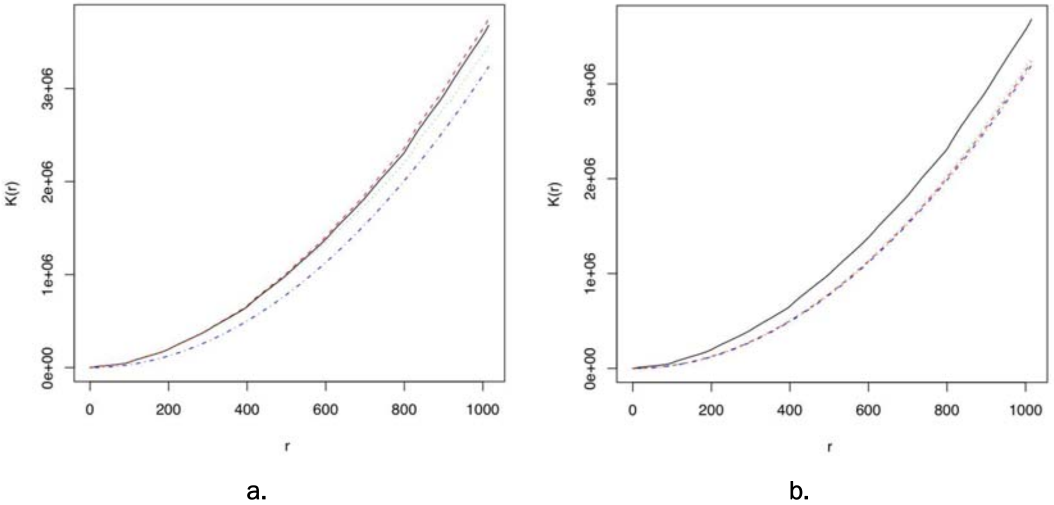


Figure 8. K-function (a) and significance envelopes (b) for homicide events in Watts.

K-functions

Using K-functions to characterize the distribution of events shows that the location of homicides follows a clustered pattern, since the calculated K-function lies above the simulated envelopes (Figure 8b).

A clustered pattern is of special interest in the analysis of crime because it suggests that crime events are not scattered randomly throughout the study area. If crime events were evenly dispersed, crime-fighting operations would likely be more costly since they would expend more police resources over more, broader areas, instead of targeting areas known to have unusually high levels of reported criminal activity. It is important to bear this in mind when assessing how the degree of clustering changes when data quality deteriorates (the objective of the last section), as it then becomes possible to formulate a close relation between data quality and the costs and, potentially, the effectiveness of those operations.

4 Assessing Data Quality

This chapter documents an exercise to simulate a process data quality deterioration. This exercise was meant to evaluate the impact on spatial analysis statistics of having to deal with bad data quality. Some simulation runs were performed using *spdep*[*] (Bivand 2008) for the aggregated data level. and (for the case of point data level) using *Spatstat* (Baddeley and Turner 2005). Both are open source modules that run under R.

The empirical application began with the dataset for the homicides in Watts, CA, and assumed this is the benchmark with the desired data quality. Two processes were then simulated: one to introduce bad quality in the form of underreporting, and a second in the form of precision errors in the location of the events.

Sensitivity to underreporting

It is very difficult to know exactly the degree of data quality in a sample dataset. In some fields, the data can be cross-validated with other sources to evaluate its quality. For example, in environmental studies, the location of precision in the location of forests or plants can be cross-validated with aerial photographs (Köl et al. 1999). In this case, there is no usual way to cross-validate, which is why simulations were used to illustrate how clusters and clustering change when data is underreported

Homogeneous underreporting

In the case of homogeneous underreporting, the simulated process consisted in having a series of points taken out of the sample within a given radius of each point as in a thinning process. It is called “homogeneous” because all points have the same probability of being selected. The algorithm starts by selecting a given point and searching nearby points within a radius *h*, then sequentially eliminating points that lie within the search radius until each point does not have a neighboring event within the search radius. The thinning process is done sequentially so that, for instance, if two points satisfy the requirement of being located within *h*, just one is eliminated.

[*] spdep: Spatial dependence: weighting schemes, statistics and models. for details, see: http://cran.r-project.org/web/packages/spdep/index.html

Figure 9 shows the candidate points to be eliminated for a radius *h* are A and B. A non-sequential thinning will result in both points being eliminated as in Figure 9b, whereas a sequential thinning will leave one of the points as shown Figure 9c. This simulated process results in a quick reduction in the peaks of the distribution as given by the kernel densities (Figure 10). K-functions also show reductions in the degree of clustering as the radius of the homogeneous thinning increases (Figure 11).

Heterogeneous underreporting

Heterogeneous underreporting was simulated using a heterogeneous thinning process. In this process, there is a function that creates a value for each point that is inversely related to the population density in each census block containing point. Because it would be undesirable for all the points from a given census block to disappear, then a uniform distribution was generated that is compared to the normalized values of the population density, and that drops a point if the normalized values are below the uniform random variable generated.

The simulation results show that the K–function shows significant instability, especially for the larger values of *r*, the distances evaluated. At the small scale, the K–function remains relatively stable, is a sign that, for heterogeneous underreporting, the clustering is not significantly affected.

Sensitivity to geographic location precision

This section evaluates another way to feature data quality deterioration, except that it focuses on problems of precision of the coordinates where the events occur. Quality deterioration was simulated by randomly shifting the location of the homicide events.

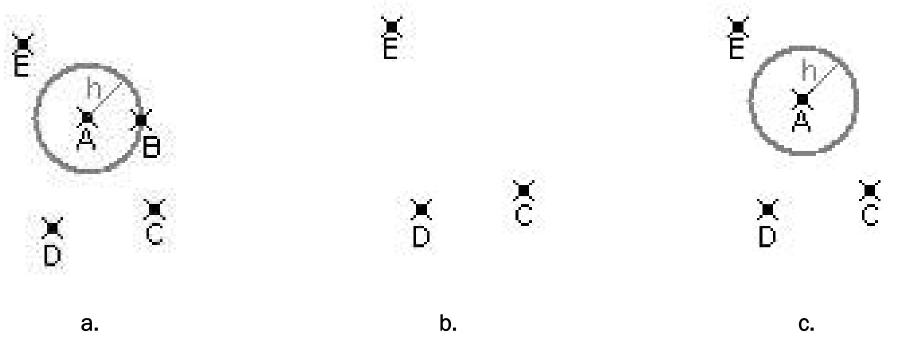
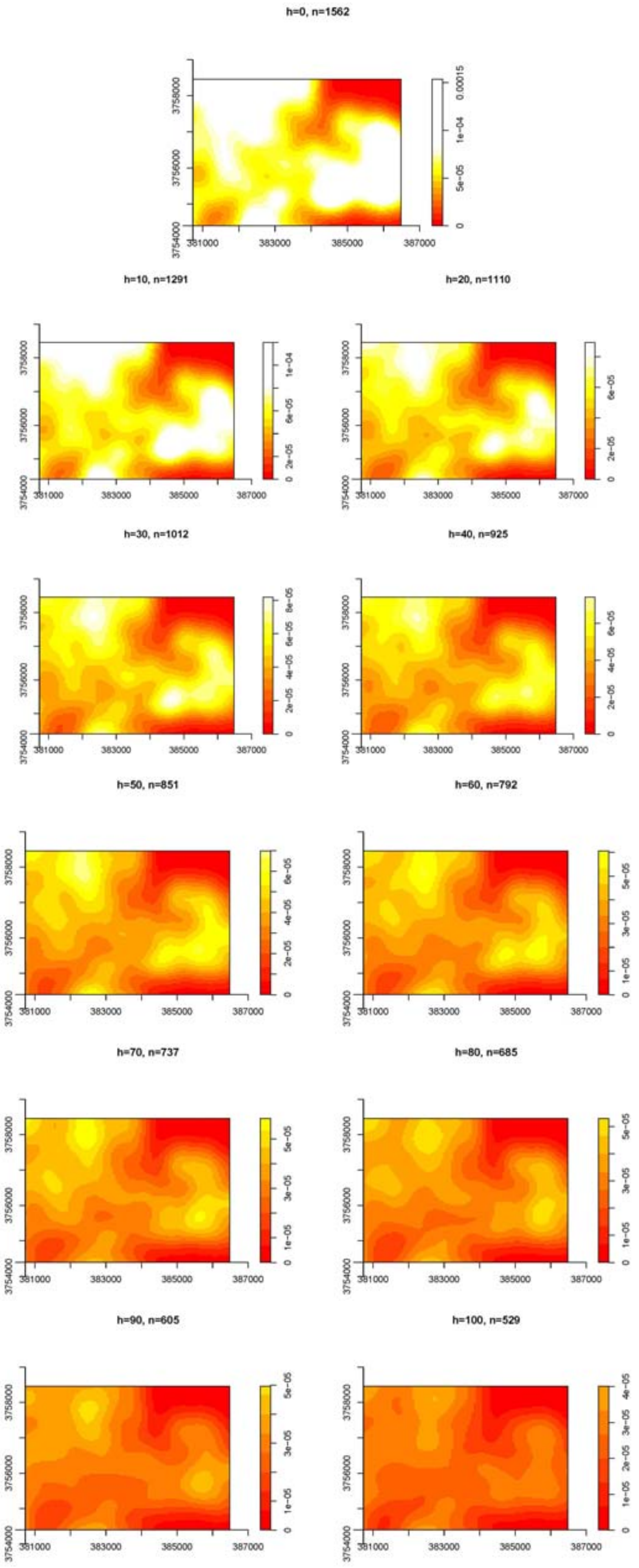


Figure 9. Homogeneous thinning: (a) candidates points to eliminate; (b) nonsequential thinning; (c) sequential thinning.


h=0, n=1562

h=10, n=1291

h=20, n=1110

h=30, n=1012

h=40, n=925

h=50, n=851

h=60, n=792

h=70, n=737

h=80, n=685

h=90, n=605

h=100, n=529

Figure 10. Kernel densities for homogeneous thinning.

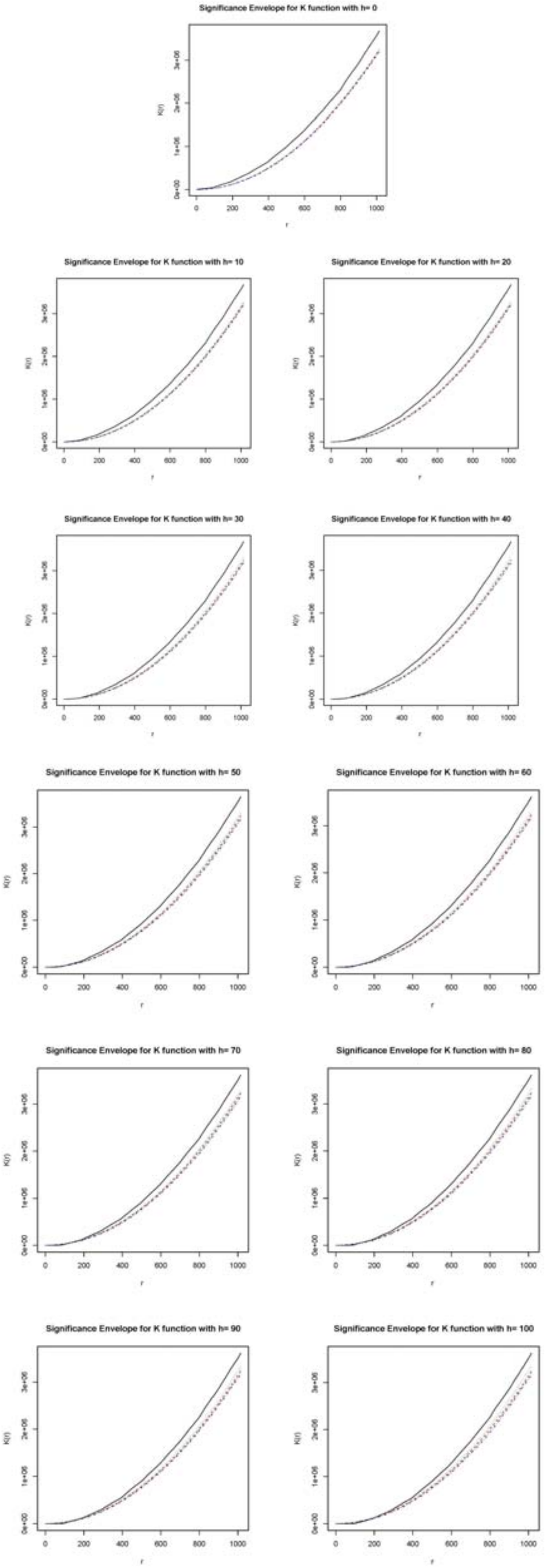


Figure 11. K-functions and envelopes for homogeneous thinning.

The procedure that performed the simulations consisted on drawing a circle around each point and randomly shifting the points within that circle. Formally, the procedure to simulate the locations was conducted as follows:

- First, a coordinate X* was calculated by randomly shifting the coordinate *X* by a fraction λi1 of the targeted radius:

$$X_i^* = X_i \pm \lambda_{i1} * h \qquad \text{Eq 6}$$

- Then the complement in the y–axis is calculated to complete a shift equivalent to a radius *h* as follows:

$$Y_i^* = Y_i \pm \lambda_{i2} * \left(\sqrt{h^2 - \left(X_i^* - X_i \right)^2} \right) \qquad \text{Eq 7}$$

Because the precision in the location of each point is not necessarily constant throughout the whole study area, i.e., that all the points are not shifted by the same magnitude *h*, a term λi2 was added that guarantees that the displacement in each point ranges from zero to the value *h*. In this case, the most notable changes occur at the aggregate level. For that reason, the following sections focus main on the aggregate level.

Global clustering

At the global level, random displacements of the events result in a strengthened degree of clustering. Figure 12 shows that the simulated plots do not provide evidence of clustering (the p-value for Moran's I is 0.12), but as the points are relocated, the clustering signs start emerging from the simulations, which reach p-values of 0.01.

The simulated process tends to locate some points in polygons nearby and, when there is a high intensity of points in a given polygon, those points are then assigned to the neighbor polygons in a particular way that results in a clustered pattern. This clustered pattern is created as a result of an artificial tendency to redistribute events from a polygon with high intensity to its neighbors, which in turn may end up having high intensity as well.

For the homicide events, this result is not surprising as the majority of events are registered along the main streets, which in turn are the ones that correspond to the polygon boundaries at census blocks or tracts. Because those events are located near the streets, a small displacement of the event locations may result in the relocation of events from one given block to its neighbors.

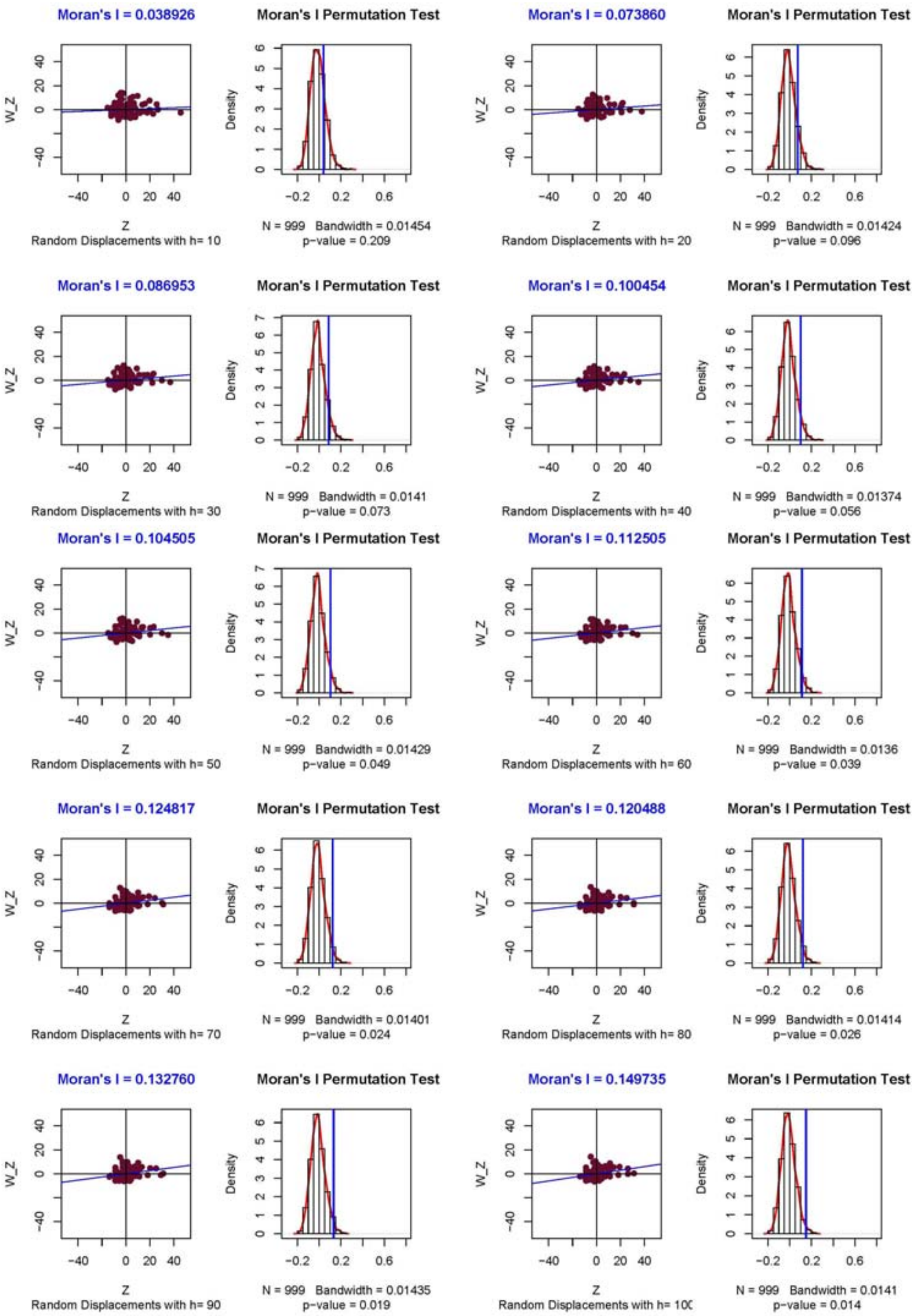


Figure 12. Global clustering with homogeneous thinning.

Local clustering

In the case of local cluster measures, the problems of precision create erroneous patterns that can mislead the conclusions regarding where the hot–spots are located and that can, in turn, erroneously target zones for operational activities.

The main conclusion from this section of the analysis is that the core of the clusters remains identified as a significant cluster when its p-value is small, e.g., 0.001 or below, but the polygons that are part of the cluster at 5 percent are not kept consistently as such (Figure 13). Another salient feature is that some polygons appear to be part of a cluster at 5 percent, but those polygons were not identified as clusters in the original dataset previous to the deterioration in quality.

In sum, only the clusters that are highly significant can be considered as such with certain degree of confidence when the quality of the data in terms of its location precision is questionable. The results of this work are consistent with those of Armstrong et al. (1999), who conducted a similar exercise aimed at developing a methodology to mask geographic data to allow researchers to have access to micro-level data without disclosing the actual location of the individuals. Using Humberside's dataset, Armstrong conducted an experiment generating certain levels of perturbation to the original coordinates and, for each level of perturbation, they simulated 500 random patterns. The authors showed that, for small levels of perturbation, the clustered pattern is still revealed, but for higher levels of perturbation, it disappears. However, when aggregation was conducted, the results of the statistical tests to detect clusters were misleading.

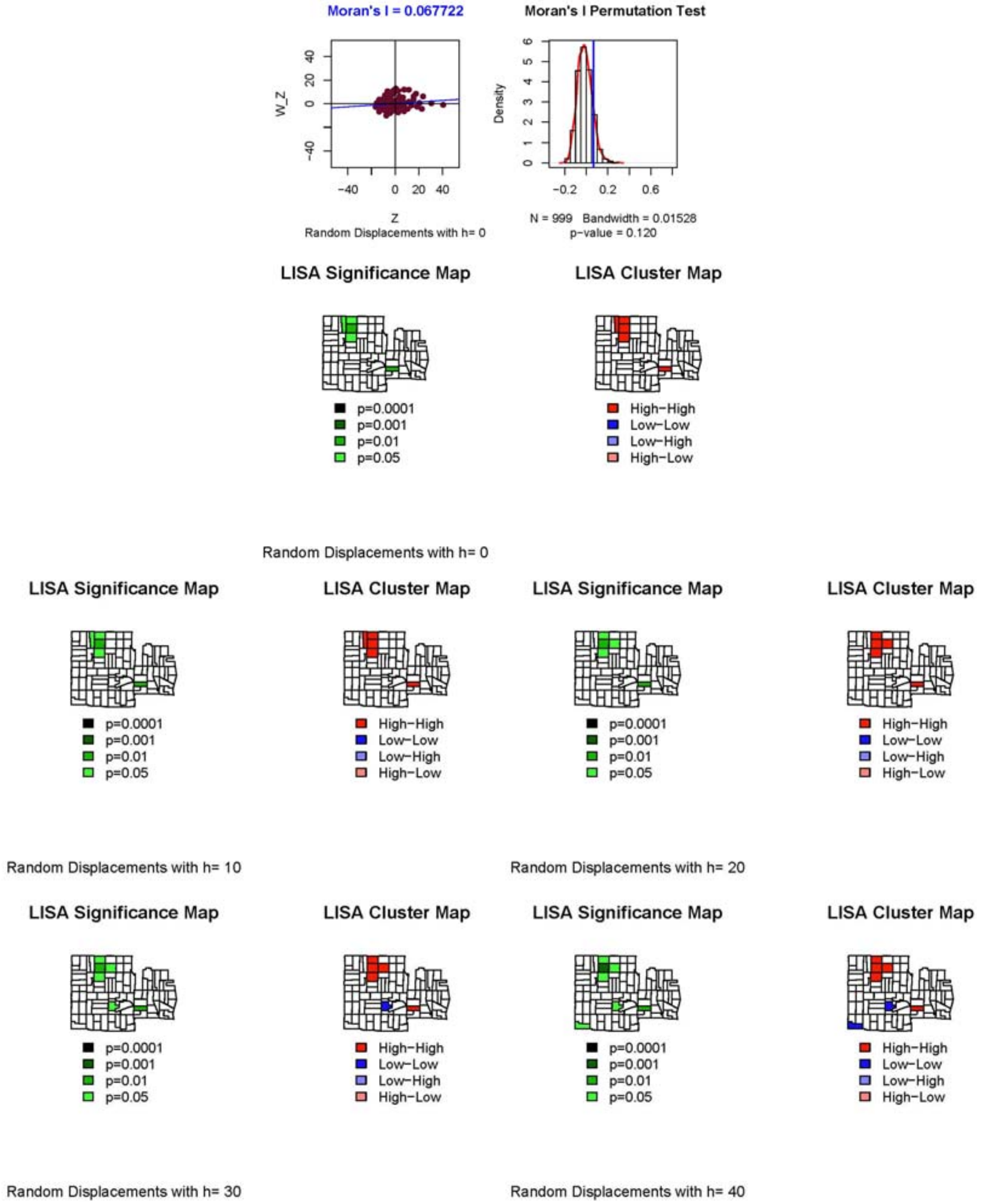
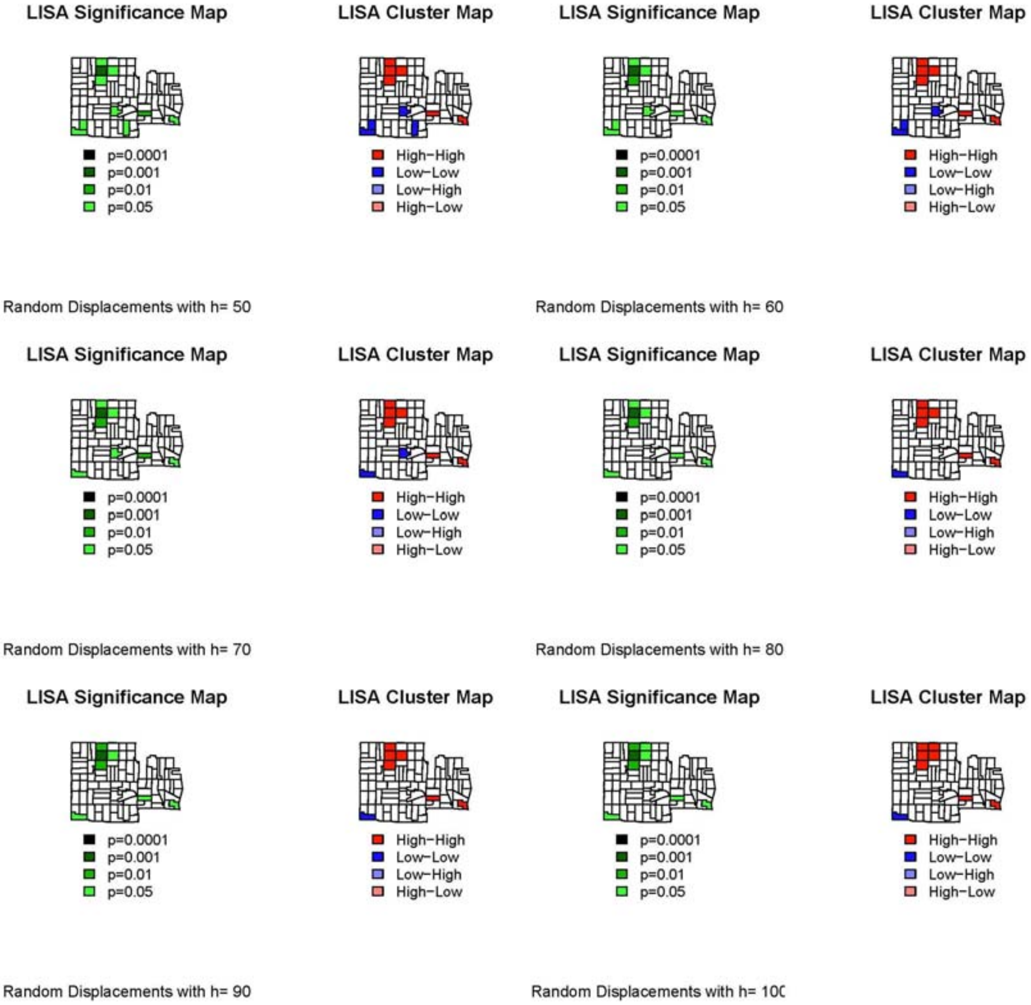

Moran's I = 0.067722
W_Z
40
20
0
-40
-40
0
20
40
Z
Random Displacements with h= 0
Moran's I Permutation Test
Density
6
5
4
3
2
1
0
-0.2
0.2
0.6
N = 999 Bandwidth = 0.01528
p-value = 0.120
LISA Significance Map
LISA Cluster Map
p=0.0001
p=0.001
p=0.01
p=0.05
High-High
Low-Low
Low-High
High-Low
Random Displacements with h= 0
Random Displacements with h= 10
Random Displacements with h= 20
Random Displacements with h= 30
Random Displacements with h= 40

Figure 13. Local clusters with homogeneous thinning.



Random Displacements with h= 50

Random Displacements with h= 60

Random Displacements with h= 70

Random Displacements with h= 80

Random Displacements with h= 90

Random Displacements with h= 10C

Figure 13. (Cont'd).

5 Conclusion

This work characterized data quality in point pattern analysis and how decreasing data quality can effect results, which may mislead decision makers into selecting the wrong COA. In the austere environments that the military often faces in which the customary protocols that ensure data collection quality cannot often be rigorously applied and where the location is not accurately characterized, analysts need to be prepared for the range of error possible in the results. Point pattern analysis of events occurring in defined geographical areas often must account for areas that are likely to be characterized by incomplete geographic data, lack of detailed disaggregated information, or information that is either outdated or being supplied by unreliable second and third parties.

This work simulated data deterioration associated with homogenous and heterogeneous underreporting, and with data imprecision, and evaluated the impact of bad data quality on spatial analysis statistics. This work concluded that only the clusters that are highly significant can be considered as such with certain degree of confidence when the quality of the data in terms of its location precision is questionable. For small levels of perturbation, clustered patterns were still revealed, but for higher levels of perturbation, they disappear.

Acronyms and Abbreviations

Term	Spellout
ACUSTO	Actionable Cultural Understanding for Support to Tactical Operations
ANSI	American National Standards Institute
AO	area of operation
ASAALT	Assistant Secretary of the Army for Acquisition, Logistics, and Technology
CERL	Construction Engineering Research Laboratory
CSR	complete spatial randomness
ERDC	Engineer Research and Development Center
GIS	geographic information system
IPB	Intelligence Preparation of the Battlefield
LISA	Local Indicator of Spatial Association
MDMP	military decision making process
NSN	National Supply Number
OE	operational environment
OMB	Office of Management and Budget
PO	Post Office
STAC	Spatial and Temporal Analysis of Crime package
TR	Technical Report
URL	Universal Resource Locator
WWW	World Wide Web

References

Anselin, L. 1995. Local Indicators of Spatial Association: LISA. *Geographical Analysis* 27(2):93–115.

Armstrong, M., G. Rushton, and D. Zimmerman. 1999. Geographically masking health data to preserve confidentiality. *Stat Med* 18:497–525.

Baddeley, A., and R. Turner. 2005. Spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software* 12(6):1–42.

Bates, S. 1987. Spatial and temporal analysis of crime. *Research Bulletin*. Chicago: Illinois Criminal Justice Information Authority. April 1987.

Bichler-Robertson, G., and M. Johnson. 2001. *Connecting Environmental Cues to Commercial Burglary Concentrations: Combining Theory and Practice into a Blended Approach*, http://cjrc.csusb.edu/CPAL/projects.html

Bivand, R. 2008. with contributions by Luc Anselin, Olaf Berke, Rew Bernat, Marilia Carvalho, Yongwan Chun, Carsten Dormann, Stéphane Dray, Rein Halbersma, Nicholas Lewin-Koh, Jielai Ma, Giovanni Millo, Werner Mueller, Hisaji Ono, Pedro Peres-Neto, Markus Reder, Michael Tiefelsdorf, Danlin Yu. *spdep: Spatial Dependence: Weighting Schemes, Statistics and Models. Package Version 0.4-17.*

Block, C. 1995. STAC hot-spot areas: a statistical tool for law enforcement decisions. *Crime Analysis Through Computer Mapping*.

Köl, M., J. Schnellbächer, and A. Grünig. 1999. From data accuracy to data quality: Using spatial statistics to predict the implications of spatial error in point data. In *Quantifying Spatial Uncertainty in Natural Resources: Theory and Applications for GIS and Remote Sensing*, Ann Arbor, Michigan.

Levine, N. 2004. *CrimeStat III: A Spatial Statistics Program for the Analysis of Crime Incident Locations*. Ned Levine & Associates, Houston, TX, and the National Institute of Justice, Washington, DC. November.

Levine, N. 2005. Crime mapping and the CrimeStat program. *Geographical Analysis* 38(1):41–56.

Ratcliffe, J. H. 2004. The hotspot matrix: A framework for spatio-temporal targeting of crime reduction. *Police Practice and Research* 5(1): 05-23.

Reichman, J. 2008. *Applying GIS in Stability Operations and Support Operations (SOSO)*. Unpublished manuscript: The National Geospatial Intelligence Agency.

Satyanarayana, P., and S. Yogendran. 2009. *Military Applications of GIS. IIC Technologies Private Limited, Hyderabad, India*, Unpublished Manuscript, gisdevelopment.net/application/military/overview/militaryf0002pf.htm

Ripley, B. 1976. The second-order analysis of stationary point processes. *Journal of Applied Probability* 13(2):255–266.

Waller, L., and C. Gotway. 2004. *Applied Spatial Analysis of Public Health Data*. John Wiley, Hoboken, NJ.

Williamson, D., S. McLafferty, P. McGuire, and T. Ross. 2001. Tools in the spatial analysis of crime. In A. Hirschfield, and K. Bowers, eds., *Mapping and Analyzing Crime Data: Lessons from Research and Practice*. New York: Taylor and Francis London, 187–203.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE (DD-MM-YYYY)
25-06-2009

2. REPORT TYPE
Final

3. DATES COVERED (From - To)

4. TITLE AND SUBTITLE
Actionable Cultural Understanding for Support to Tactical Operations (ACUSTO):
The Effect of Data Quality on Spatial Analysis Results

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT

6. AUTHOR(S)
Luis Galvis, Patrick J. Guertin, and William D. Meyer

5d. PROJECT NUMBER
622784AT41

5e. TASK NUMBER

5f. WORK UNIT NUMBER
21 2040

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Engineer Research and Development Center (ERDC)
Construction Engineering Research Laboratory (CERL)
PO Box 9005,
Champaign, IL 61826-9005

8. PERFORMING ORGANIZATION REPORT NUMBER
ERDC/CERL TR-09-15

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Assistant Secretary of the Army for Acquisition, Logistics, and Technology (ASAALT)
2511 Jefferson Davis Highway, Presidential Towers
Arlington, VA 22202-3911

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Developing cultural information into cultural knowledge for military operations is predominantly an intelligence activity that takes place within the Military Decision Making Process. The products of such efforts are routinely classified and unusable by the tactical war fighter. The Actionable Cultural Understanding for Support to Tactical Operations (ACUSTO) research effort was undertaken to provide a product for enhanced cultural understanding that will be accessible to the tactical war fighter. This is done by combining spatial and explicit content analysis of open source news media to provide cultural understanding in the operational environment that can be disseminated down to the lowest tactical level. The development of actionable intelligence for counterinsurgency parallels the study of civilian criminal events (widely covered in open source media) and can exploit the methodological approaches that emphasize spatially explicit information. Crime research is conducted at aggregate levels, which implies the aggregation of a series of points representing events to areas representing higher scales. This work focused on data quality in point pattern analysis, and on the effect that different levels of data accuracy and precision have on policy recommendations.

15. SUBJECT TERMS

spatial analysis, ACUSTO, cultural knowledge, data management

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	SAR	40	**19b. TELEPHONE NUMBER (include area code)**